AD-770 130

SOME TOPICS IN STATISTICAL INFORMATION THEORY

S. Kullback

George Washington University

Prepared for:

Office of Naval Research

15 November 1973

DISTRIBUTED BY:

NTIS

National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

UNCLASSIFIED
Security Classification

AD770 130

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) THE GEORGE WASHINGTON UNIVERSITY DEPARTMENT OF STATISTICS WASHINGTON, D.C. 20006	2a. REPORT SECURITY CLASSIFICATION 2b. GROUP

3. REPORT TITLE

SOME TOPICS IN STATISTICAL INFORMATION THEORY

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

TECHNICAL REPORT

5. AUTHOR(S) (First name, middle initial, last name)

S. KULLBACK

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
November 15, 1973	26	15

8a. CONTRACT OR GRANT NO N00014-67-A-0214-0015 b. PROJECT NO NR-042-267 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 13 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10. DISTRIBUTION STATEMENT

Unlimited. Reproduction in whole or in part is permitted for any purpose of the United States Government.

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
	OFFICE OF NAVAL RESEARCH STATISTICS & PROBABILITY PROGRAM ARLINGTON, VIRGINIA 22217

13. ABSTRACT

Attention is focused on informational properties of sub-sigma-algebras of the fundamental probability space in contrast to the discussion in Information Theory and Statistics (Wiley 1959, Dover Publications Inc. 1968) where attention is devoted to informational properties of statistics, that is, random variables. Properties of relative conditional expectations and results on separable sigma-algebras are stated for later use. The discrimination information in a sigma-algebra (sub-sigma-algebra) is defined and various properties developed, in particular the relation between information and sufficiency. The integral representative theorem for discrimination information is derived by methods believed to be more inherently information-theoretic than others that have been presented. Monotonic properties of conditional discrimination information are derived with respect to either the conditioning sub-sigma-algebra or the conditioned sub-sigma-algebra.

32

DD FORM 1473 (PAGE 1)
1 NOV 65

S/N 0101-807-6801

UNCLASSIFIED
Security Classification

UNCLASSIFIED
Security Classification

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
STATISTICAL INFORMATION THEORY						
MINIMUM DISCRIMINATION						
SUB-SIGMA-ALGEBIAS						

ii

DD FORM 1 NOV 65 1473 (BACK)
(PAGE 2)

UNCLASSIFIED
Security Classification

SOME TOPICS IN STATISTICAL INFORMATION THEORY

by

S. KULLBACK

TECHNICAL REPORT NO. 13

November 15, 1973

PREPARED UNDER CONTRACT N00014-67-A-0214-0015

(NR-042-267)

OFFICE OF NAVAL RESEARCH

Herbert Solomon, Project Director

DEPARTMENT OF STATISTICS

THE GEORGE WASHINGTON UNIVERSITY

WASHINGTON, D.C. 20006

SOME TOPICS IN STATISTICAL INFORMATION THEORY

by

S. Kullback

Summary

Attention is focused on informational properties of sub-sigma-algebras of the fundamental probability space in contrast to the discussion in Information Theory and Statistics where attention is devoted to informational properties of statistics that is, random variables. In particular, the integral representation theorem for discrimination information is derived by methods believed to be more inherently information-theoretic than others that have been presented. Monotonic properties of conditional discrimination information are derived.

0. Preliminaries. In [11] attention was devoted to informational properties of statistics, that is, random variables. In this exposition however, the discussion deals with informational properties of sub-sigma-algebras of the fundamental probability space. In particular, we shall present a proof of the integral representation theorem of discrimination information which is believed to be more information-theoretic in approach than other

derivations of this basic result.

We present here certain notations, lemmas, and results on separable sigma-algebras which we shall use in this exposition.

We shall operate in the probability space (Ω, A, P). Let $Z_t(\omega)$ and $Z(\omega)$ be non-negative random variables such that

(0.1) $\mu_t(A) = \int_A Z_t(\omega)dP(\omega), \quad \mu(A) = \int_A Z(\omega)dP(\omega), \; A \in A$

are probability measures. We also write (0.1) in the Radon-Nikodym differential formalism as

(0.2) $d\mu_t = Z_t dP, \quad d\mu = ZdP.$

$Z_t(\omega)$ and $Z(\omega)$ may be considered as generalized probability densities. If we assume that μ_t is absolutely continuous with respect to μ, that is, $\mu_t << \mu$, then

(0.3) $d\mu_t = W_t d\mu = W_t ZdP = Z_t dP, \quad W_t = Z_t/Z$ a.s.

so that W_t is a likelihood ratio. We shall also require sequences of the generalized densities, corresponding probability measures, and likelihood ratios, that is,

$$d\mu_{t_n} = Z_{t_n} dP, \quad d\mu_n = Z_n dP, \quad n = 1,2,...$$

(0.4) $d\mu_{t_n} = W_{t_n} d\mu_n = W_{t_n} Z_n dP = Z_{t_n} dP, \; W_{t_n} = Z_{t_n}/Z_n$ a.s.

We shall have occasion to deal with the properties of relative conditional expectations as described in [13, p. 344].

Let B be a sub-sigma-algebra of the sigma-algebra A, that is, $B \subset A$. Corresponding to (0.2) and (0.3)

$$(0.5)\quad d\mu_{\mathcal{B}} = E^{\mathcal{B}}Z\,dP_{\mathcal{B}},\ d\mu_{t\,\mathcal{B}} = E^{\mathcal{B}}Z_t\,dP_{\mathcal{B}},\ d\mu_{t\,\mathcal{B}} = E_Z^{\mathcal{B}}W_t\,d\mu_{\mathcal{B}}$$

$$E_Z^{\mathcal{B}}W_t = E_Z^{\mathcal{B}}\frac{Z_t}{Z} = (E^{\mathcal{B}}Z\cdot\frac{Z_t}{Z})/E^{\mathcal{B}}Z = E^{\mathcal{B}}Z_t/E^{\mathcal{B}}Z,$$

where $P_{\mathcal{B}}$ is the restriction of P to $\mathcal{B}$ defined by $P_{\mathcal{B}}(B)=P(B)$, $B\in\mathcal{B}$, and

$$(0.6)\quad \mu_{\mathcal{B}}(B) = \int_B Z\,dP = \int_B (E^{\mathcal{B}}Z)\,dP_{\mathcal{B}},\quad B\in\mathcal{B}$$

$$(0.7)\quad \int_B (E_Z^{\mathcal{B}}X)\,d\mu_{\mathcal{B}} = \int_B X\,d\mu,\quad B\in\mathcal{B}$$

$$(0.8)\quad E^{\mathcal{B}}ZX = E^{\mathcal{B}}Z\cdot E_Z^{\mathcal{B}}X\ .$$

We shall need two results from probability theory (see for example, [13, p. 140, prob. 16, 17] which we state as lemmas.

<u>Lemma 0.1</u>. $\int|Z_{t_n} - Z_t|dP \to 0$, resp.

$\int|Z_n - Z|dP \to 0$ as $n \to \infty$, if and only if

$\int_A Z_{t_n}dP \to \int_A Z_t\,dP$, resp. $\int_A Z_n\,dP \to \int_A Z\,dP$ as $n \to \infty$

uniformly in $A\in\mathcal{A}$.

<u>Lemma 0.2</u>. If $Z_{t_n} \xrightarrow{P} Z_t$ resp. $Z_n \xrightarrow{P} Z$, then $\int_A Z_{t_n}dP \to \int_A Z_t\,dP$ resp. $\int_A Z_n\,dP \to \int_A Z\,dP$ as $n \to \infty$ uniformly in $A\in\mathcal{A}$. The convergence in probability may be replaced by almost sure convergence.

Note that Lemmas 0.1 and 0.2 provide the chain of implications

$$(0.9)\quad \mu_n(A) \to \mu(A),\ \text{uniformly in } A\in\mathcal{A} \Leftrightarrow Z_n \xrightarrow{L_1} Z \Rightarrow$$

$$\Rightarrow Z_n \xrightarrow{P} Z \Rightarrow \mu_n(A) \to \mu(A),\ \text{uniformly in } A\in\mathcal{A} \text{ and a similar}$$

one with the subscript t, where

(0.10) $\quad Z_n \xrightarrow{L_1} Z \Leftrightarrow \int |Z_n - Z| dP \to 0.$

We assemble here certain results on separable sigma-algebras which we shall need [13].

(0.11) A separable sigma-algebra is a sigma-algebra that is generated by (is minimal over) a countable class of sets.

(0.12) The minimal sigma-algebra over the union of a countable class of separable sigma-algebras is also a separable sigma-algebra.

(0.13) The Borel sigma-algebra on the real line is separable.

(0.14) The inverse image of a separable sigma-algebra by a measurable transformation is a separable sigma-algebra.

(0.15) The sub-sigma-algebra induced by a random variable or a countable class of random variables is a separable sigma-algebra.

(0.16) A finite (countable) partition of a space Ω is a finite (countable) sequence of sets A_i such that

$$\sum A_i = \Omega \qquad A_i \cap A_j = \emptyset \quad i \neq j.$$

If $\mathcal{A}$ is a sigma-algebra of subsets of Ω then the partition is measurable $\mathcal{A}$ if $A_i \in \mathcal{A}$ for all i. Let $\mathcal{E} = \{E_i\}$ be an $\mathcal{A}$-measurable partition. The $\mathcal{A}$-measurable partition $\mathcal{D}=\{D_j\}$ is said to be a subpartition of $\mathcal{E}$ or finer than the partition $\mathcal{E}$ if each $D_j \in \mathcal{D}$ is such that $D_j \subset E_i \in \mathcal{E}$ and we denote this by $\mathcal{D} < \mathcal{E}$ or $\mathcal{E} > \mathcal{D}$.

(0.17) A sequence of partitions $\{E^n\}$ is said to be regular if each E^n is a finite partition and

$$E^1 > E^2 > E^3 > \ldots \quad .$$

(0.18) Let $\underline{E}^n$ denote the finite algebra generated by the partition E^n, then corresponding to (0.17)

$$\underline{E}^1 \subset \underline{E}^2 \subset \underline{E}^3 \subset \ldots$$

and $\bigcup_1^\infty \underline{E}^n$ is an algebra. Let $\underline{E}$ be the minimal sigma-algebra over $\bigcup_1^\infty \underline{E}^n$, then $\underline{E}$ is said to be generated by the regular sequence of partitions (0.17)

(0.19) It is clear that if a sigma-algebra is generated by a regular sequence of partitions then it is separable. Conversely, every separable sigma-algebra Λ can be generated by a regular sequence of partitions.

1. Introduction. The result in Corollary 3.2, page 16 [11] suggests that the discrimination information in the sub-sigma-algebra $\mathcal{B} \subset \mathcal{A}$ generated by the partition $\{B_i\}$, $i = 1,2,\ldots,n$, $B_i \in \mathcal{A}$, $\sum_{i=1}^{n} B_i = \Omega$, be defined by (we shall use natural logarithms)

$$(1.1) \quad I(\mathcal{B};\mu_t,\mu) = \sum_{i=1}^{n} \mu_t(B_i)\,\ell n \frac{\mu_t(B_i)}{\mu(B_i)}, \quad B_i \in \mathcal{B} \subset \mathcal{A}.$$

Because of the convexity of the function $x \log \frac{x}{y}$ for non-negative x and y, and additivity of the measures for disjoint sets, for $A_1, A_2 \in \mathcal{A}$, $A_1 \cap A_2 = \emptyset$,

$$(1.2) \quad \mu_t(A_1)\,\ell n \frac{\mu_t(A_1)}{\mu(A_1)} + \mu_t(A_2)\,\ell n \frac{\mu_t(A_2)}{\mu(A_2)}$$
$$\geq (\mu_t(A_1) + \mu_t(A_2))\,\ell n \frac{\mu_t(A_1)+\mu_t(A_2)}{\mu(A_1)+\mu(A_2)}$$
$$= \mu_t(A_1+A_2)\,\ell n \frac{\mu_t(A_1+A_2)}{\mu(A_1+A_2)}.$$

The property in (1.2) suggests that the discrimination information in $\mathcal{A}$ be defined by (cf. [1],[2],[3],[6],[7],[8],[10],[14],[15])

$$(1.3) \quad \bar{I}(\mathcal{A};\mu_t,\mu) = \sup_{A_i \in \mathcal{A}} \sum \mu_t(A_i)\,\ell n \frac{\mu_t(A_i)}{\mu(A_i)}$$

where the sup is taken over all possible $\mathcal{A}$-measurable finite partitions of Ω. For convenience hereafter we shall omit the μ_t and μ in $I(\mathcal{B};\mu_t,\mu)$ and $I(\mathcal{A};\mu_t,\mu)$ unless needed for clarification.

If μ_t is not absolutely continuous with respect to μ,

that is, there exists an $A \in A$ such that $\mu(A) = 0$, $\mu_t(A) \neq 0$ then $\bar{I}(A) = \infty$. Accordingly we shall assume that $\mu_t << \mu$. Note that $\bar{I}(A)$ may be infinite in this case also. $\mu_t << \mu$ is a necessary condition for $\bar{I}(A) < \infty$. (See [11, pp. 5, 6. Prob 5.7, p. 10]

It also seems intuitively reasonable to have defined the discrimination information in A by (cf. [11, p. 5])

$$(1.4) \quad I(A) = \int Z_t(\omega) \ln \frac{Z_t(\omega)}{Z(\omega)} dP(\omega).$$

The integral representation (1.4) may also be written as

$$(1.5) \quad I(A) = \int Z_t \ln \frac{Z_t}{Z} dP = \int W_t \ln W_t d\mu = \int (\ln W_t) d\mu_t = \int d\mu_t \ln \frac{d\mu_t}{d\mu} .$$

Using the additivity of the integral and Jensen's inequality for convex functions [11, p. 16]

$$(1.6) \quad \int W_t \ln W_t d\mu = \sum_{i=1}^{n} \int_{A_i} W_t \ln W_t d\mu, \quad A_i \in A, \quad \sum_{i=1}^{n} A_i = \Omega,$$

$$(1.7) \quad \int_{A_i} W_t \ln W_t d\mu \geq \int_{A_i} W_t d\mu \ln \frac{\int_{A_i} W_t d\mu}{\int_{A_i} d\mu} = \mu_t(A_i) \ln \frac{\mu_t(A_i)}{\mu(A_i)},$$

$$(1.8) \quad \int W_t \ln W_t d\mu \geq \sum_{i=1}^{n} \mu_t(A_i) \ln \frac{\mu_t(A_i)}{\mu(A_i)},$$

hence

$$(1.9) \quad I(A) = \int W_t \ln W_t d\mu \geq \sup_{A_i \in A} \sum \mu_t(A_i) \ln \frac{\mu_t(A_i)}{\mu(A_i)} = \bar{I}(A),$$

where the sup is taken over all possible A-measurable finite partitions of Ω. A proof of the reverse of the inequality in (1.9) may be obtained following the method in [15, pp. 24-25]

or that in [8]. We shall not pursue this matter any further at this point but state the integral representation theorem

<u>Theorem 1.1</u>. $\bar{I}(A) = I(A)$, that is (1.3) and (1.5) define the same value of the discrimination information.

Proofs of this theorem have involved martingale theory (e.g. [10]) or the use in [8] of the convexity property in conjunction with the Darboux-Young approach to the integral [13, p. 143 Ex 29].

The proof to be presented later in this exposition is believed to be intrinsically more information-theoretic in nature. For other approaches see [2],[7],[14],[15].

Note that if in the probability space (Ω,A,P), A is generated by a finite partition $\{A_i\}$, then (0.1) is

(1.10) $\mu_t(A_i) = Z_t(\omega)P(A_i)$, $\mu(A_i) = Z(\omega)P(A_i)$, $\omega\in A_i$

and both (1.3) and (1.5) yield

(1.11) $\bar{I}(A) = \sum_{i=1}^{n} \mu_t(A_i)\ \ell n \dfrac{\mu_t(A_i)}{\mu(A_i)} = I(A).$

We remark that instead of starting with the probability space (Ω,A,P) we could have started with the measure space (Ω,A,λ) where λ is a sigma-finite measure on A. We assume the existence of the non-negative A-measurable function $X(\omega)$ such that

(1.12) $P(A) = \int_A X(\omega)d\lambda(\omega)$, $A\in A$

is a probability measure on A. We then have (see [11, p. 5])

(1.13) $d\mu_t = Z_t\, dP = Z_t X d\lambda = f_t\, d\lambda, \quad d\mu = Z dP = ZX d\lambda = f d\lambda\, [\lambda]$

(1.14) $\int Z_t \, \ell n \frac{Z_t}{Z}\, dP = \int Z_t \, \ell n \frac{Z_t}{Z}\, X d\lambda = \int f_t \, \ell n \frac{f_t}{f}\, d\lambda.$

2. Discrimination Information in a Sub-sigma-algebra. The following discussion essentially extends some of the presentation in [11, pages 1-78]. We shall use the notation in [13] and in particular properties developed in [13, Chapter VII, Conditioning, pp. 337 ff]. In particular we shall use the fact that Jensen's inequality holds a.s. for conditional expectation also. Let $\mathcal{B}$ be a sub-sigma-algebra of the sigma-algebra $\mathcal{A}$. The basic inequalit is that if g is a convex function and EX is finite, then

(2.1) $E\{g(X)\} \geq g(EX).$

In the conditional form, we have

(2.2) $E^{\mathcal{B}}\{g(X)\} \geq g(E^{\mathcal{B}}X)$ a.s.

Using the definition of conditional expectation, (2.2) and (1.4),

(2.3)
$$I(\mathcal{A}) = \int Z_t \, \ell n \frac{Z_t}{Z}\, dP = \int E^{\mathcal{B}}(Z_t \, \ell n \frac{Z_t}{Z})\, dP_{\mathcal{B}}$$
$$\geq \int E^{\mathcal{B}} Z_t \, \ell n \frac{E^{\mathcal{B}} Z_t}{E^{\mathcal{B}} Z}\, dP_{\mathcal{B}}\ .$$

We define the right-hand side of (2.3) as the discrimination information in the sub-sigma-algebra $\mathcal{B}$ and note that, using (0.5),

(2.4)
$$I(\mathcal{B}) = \int E^{\mathcal{B}} Z_t \, \ell n \frac{E^{\mathcal{B}} Z_t}{E^{\mathcal{B}} Z}\, dP_{\mathcal{B}} = \int E_Z^{\mathcal{B}} W_t \, \ell n \, E_Z^{\mathcal{B}} W_t \, d\mu_{\mathcal{B}}$$
$$= \int (\ell n \, E_Z^{\mathcal{B}} W_t)\, d\mu_{t\,\mathcal{B}} = \int d\mu_{t\,\mathcal{B}} \ell n \frac{d\mu_{t\,\mathcal{B}}}{d\mu_{\mathcal{B}}}$$
$$= \operatorname{Sup} \sum \mu_t(B_k) \, \ell n \frac{\mu_t(B_k)}{\mu(B_k)}$$

where the Sup is taken over all finite $\mathcal{B}$-measurable partitions of Ω (see (2.15) and section 1). We can now state:

Theorem 2.1. If $\mathcal{B}$ is a sub-sigma-algebra of $\mathcal{A}$ then

(2.5) $\quad I(\mathcal{A}) \geq I(\mathcal{B})$.

Note that the coarsest possible sub-sigma-algebra is that generated by $(\emptyset, \Omega)$ and denoting it by $\mathcal{B}_0$, $E^{\mathcal{B}_0} Z_t = EZ_t = 1$, $E^{\mathcal{B}_0} Z = EZ = 1$ and

(2.6) $\quad I(\mathcal{B}_0) = 0$.

Theorem 2.2. $I(\mathcal{A}) \geqq 0$ with equality if and only if $W_t = Z_t/Z = 1$ a.s. (See [11, Theorem 3.1, p. 14]).

From (1.5) and (2.4), using the result that

$$(2.7) \quad \int W_t \ \ln E_Z^{\mathcal{B}} W_t \, d\mu = \int E_Z^{\mathcal{B}} (W_t \ \ln E_Z^{\mathcal{B}} W_t) \, d\mu_{\mathcal{B}} = \int E_Z^{\mathcal{B}} W_t \ \ln E_Z^{\mathcal{B}} W_t \, d\mu_{\mathcal{B}}$$

we have

$$(2.8) \quad I(\mathcal{A}) - I(\mathcal{B}) = \int W_t \ \ln W_t \, d\mu - \int E_Z^{\mathcal{B}} W_t \ \ln E_Z^{\mathcal{B}} W_t \, d\mu_{\mathcal{B}}$$

$$= \int W_t \ \ln (W_t / E_Z^{\mathcal{B}} W_t) \, d\mu$$

$$= \int \frac{Z_t}{Z} \ \ln \frac{Z_t/Z}{E^{\mathcal{B}} Z_t / E^{\mathcal{B}} Z} \, d\mu.$$

Since

$$\int W_t \, d\mu = \int \frac{Z_t}{Z} \, d\mu = \int Z_t \, dP = 1,$$

$$(2.9) \quad \int E_Z^{\mathcal{B}} W_t \, d\mu_{\mathcal{B}} = \int (E^{\mathcal{B}} Z_t / E^{\mathcal{B}} Z) \, d\mu_{\mathcal{B}}$$

$$= \int E^{\mathcal{B}} Z_t \, dP_{\mathcal{B}} = \int Z_t \, dP = 1,$$

the right-hand side of (2.8) is a discrimination information value, and as such non-negative. Let us define, using (2.8)

above, and (0.2) and (0.5)

$$(2.10)\quad I(A|B) = \int W_t \ \ln\ (W_t/E_Z^B W_t)d\mu = \int Z_t \ \ln \frac{Z_t/E^B Z_t}{Z/E^B Z}\, dP =$$

$$= \int d\mu_t \ \ \ln \frac{d\mu_t/d\mu_{tB}}{d\mu/d\mu_B} \quad ,$$

that is, $I(A|B)$ is the conditional discrimination information in A given B and hence

Theorem 2.3. If B is a sub-sigma-algebra of A

$$(2.11)\quad I(A) = I(B) + I(A|B).$$

Theorem 2.4. If B is a sub-sigma-algebra of A, then

$$(2.12)\quad I(A) = I(B)$$

if and only if $I(A|B) = 0$, that is, if and only if

$$(2.13)\quad Z_t/Z = E^B Z_t/E^B Z. \quad \text{a.s.}$$

Proof. Apply Theorem 2.2 and (2.10).

If B is a sub-sigma-algebra of A and satisfies Theorem 2.4, then we say that B is a sufficient sub-sigma-algebra for A. (See [13, p. 346], [11, pp. 18-22].)

If B is generated by a finite partition, then on a non-null atom $B \in B$, the conditional expected value $E^B X$ is a constant and its value is

$$(2.14)\quad E^B X = \frac{1}{P(B)} \int_B X dP, \ \omega \in B.$$

Thus

$$(2.15)\quad I(B) = \int E^B Z_t \ \ \ln \frac{E^B Z_t}{E^B Z}\, dP_B = \sum \int_{B_i} E^B Z_t \ \ \ln \frac{E^B Z_t}{E^B Z}\, dP_B$$

$$= \sum_i P(B_i) \cdot \frac{1}{P(B_i)} \int_{B_i} Z_t\, dP \ \ \ln \frac{\int_{B_i} Z_t\, dP}{\int_{B_i} Z dP} = \sum_i \mu_t(B_i) \ \ \ln \frac{\mu_t(B_i)}{\mu(B_i)} \quad .$$

Note that in this case

$$(2.16)\quad \frac{Z}{E^{B}Z} = Z/\frac{1}{P(B)}\int_B Z\,dP = Z/(\mu(B)/P(B)), \quad \omega\in B$$

and

$$(2.17)\quad I(A|B) = \int Z_t \ \ell n \frac{Z_t/E^{B}Z_t}{Z/E^{B}Z}\,dP$$

$$= \sum_i \int_{B_i} Z_t \ \ell n \frac{Z_t/E^{B}Z_t}{Z/E^{B}Z}\,dP$$

$$= \sum_i \int_{B_i} Z_t \ \ell n \frac{Z_t\, P(B_i)\,\mu(B_i)}{\mu_t(B_i)\,Z P(B_i)}\,dP$$

$$= \sum_i \mu_t(B_i) \int_{B_i} \frac{Z_t}{\mu_t(B_i)} \ \ell n \frac{Z_t/\mu_t(B_i)}{Z/\mu(B_i)}\,dP$$

$$= \sum_i \mu_t(B_i)\, I(A|B_i)$$

where

$$(2.18)\quad I(A|B_i) = \int_{B_i} \frac{Z_t}{\mu_t(B_i)} \ \ell n \frac{Z_t/\mu_t(B_i)}{Z/\mu(B_i)}\,dP$$

$$= \int_{B_i} \frac{d\mu_t}{\mu_t(B_i)} \ \ell n \frac{d\mu_t/\mu_t(B_i)}{d\mu/\mu(B_i)}$$

$$= \frac{1}{\mu_t(B_i)} \int_{B_i} Z_t \ \ell n \frac{Z_t}{Z}\,dP - \ell n \frac{\mu_t(B_i)}{\mu(B_i)}$$

$$= \frac{1}{\mu_t(B_i)} \int_{B_i} W_t \ \ell n\, W_t\,d\mu - \ell n \frac{\mu_t(B_i)}{\mu(B_i)}\,.$$

Note that $I(A|B_i)$ is a discrimination information, and

$$(2.19)\quad I(A|B_i) \geq 0$$

with equality if and only if

$$(2.20) \quad Z_t/Z = \mu_t(B_i)/\mu(B_i), \quad \omega \in B_i,$$

(see [11, Corollary 3.1, p. 15].)

To obtain another representation for $I(A|\mathcal{B})$ when $\mathcal{B}$ is a separable sub-sigma-algebra of A we proceed as follows. It follows from (2.18) that

$$(2.21) \quad \sum_i \mu_t(B_i)\, I(A|B_i) = \sum_i \int_{B_i} Z_t \,\ell n \frac{Z_t}{Z}\, dP - \sum_i \mu_t(B_i)\, \ell n \frac{\mu_t(B_i)}{\mu(B_i)}$$

$$= \int Z_t \,\ell n \frac{Z_t}{Z}\, dP - \sum_i \mu_t(B_i)\, \ell n \frac{\mu_t(B_i)}{\mu(B_i)}$$

$$= I(A) - \sum_i \mu_t(B_i)\, \ell n \frac{\mu_t(B_i)}{\mu(B_i)} .$$

Now let $\Pi(\mathcal{B})$ denote the class of finite $\mathcal{B}$-measurable partitions of Ω, then (see [8])

$$(2.22) \quad \inf_{\Pi(\mathcal{B})} \sum_i \mu_t(B_i) I(A|B_i) = \inf_{\Pi(\mathcal{B})} \{I(A) - \sum_i \mu_t(B_i)\, \ell n \frac{\mu_t(B_i)}{\mu(B_i)}\}$$

$$= I(A) - \sup_{\Pi(\mathcal{B})} \sum_i \mu_t(B_i)\, \ell n \frac{\mu_t(B_i)}{\mu(B_i)}$$

$$= I(A) - I(\mathcal{B}) = I(A|\mathcal{B}).$$

From (2.22) we note that $I(\mathcal{B}|\mathcal{B}) = 0$.

Let W_t be defined as in (0.3) and denote by $\mathcal{B}_{W_t}$ the class of sets $W_t^{-1}(B)$ where B ranges over linear Borel sets, that is

(2.23) $W_t^{-1}(B) = \{\omega : W_t(\omega) \in B\}$.

$\mathcal{B}_{W_t}$ is the minimal sigma-algebra with respect to which W_t is measurable and $\mathcal{B}_{W_t}$ is a separable sub-sigma-algebra of A. For convenience we shall hereafter denote $\mathcal{B}_{W_t}$ by $\mathcal{B}_t$. We shall now show that $\mathcal{B}_t$ is a sufficient sub-sigma-algebra for A.

Theorem 2.5. $\mathcal{B}_t$ is a sufficient sub-sigma-algebra for A, that is, $I(A) = I(\mathcal{B}_t)$.

Proof.

$$(2.24)\quad I(A) = \int W_t \, \ell n\, W_t \, d\mu = \int E_Z^{\mathcal{B}_t}(W_t \, \ell n\, W_t)\, d\mu_{\mathcal{B}_t}$$

$$\geq \int E_Z^{\mathcal{B}_t} W_t \, \ell n\, E_Z^{\mathcal{B}_t} W_t \, d\mu_{\mathcal{B}_t} = I(\mathcal{B}_t) = \int W_t \, \ell n\, W_t \, d\mu = I(A)$$

where we have used the fact that since W_t is measurable $\mathcal{B}_t$ then $E_Z^{\mathcal{B}_t} W_t = W_t\,[\mu]$. Note that using (0.5)

$$E^{\mathcal{B}_t} Z_t / E^{\mathcal{B}_t} Z = Z_t / Z,$$

which is the necessary and sufficient condition (2.13) for Theorem 2.4.

3. An Information-theoretic Approach. We shall now consider an approach to the integral representation which is believed to be of interest in that it is intrinsically information-theoretic in nature.

Suppose that there is an increasing sequence of sub-sigma-algebras of $\mathcal{A}$ such that

$$(3.0) \quad \mathcal{B}_0 \subset \mathcal{B}_1 \subset \mathcal{B}_2 \subset \cdots \subset \mathcal{B}_n \subset \cdots \subset \mathcal{B} \subset \mathcal{A}$$

where $\mathcal{B}$ is the minimal sigma-algebra containing $\bigcup_n \mathcal{B}_n$, usually denoted by $\mathcal{B}_n \uparrow \mathcal{B}$, or $\mathcal{B} = \vee \mathcal{B}_n$.

Consider

$$(3.1) \quad \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_n}Z}\, dP = \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z}\, dP + \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_{n+1}}Z}\, dP$$

so that since

$$(3.2) \quad \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z}\, dP = \int E^{\mathcal{B}_{n+1}} \left(E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z}\right) dP$$

$$= \int E^{\mathcal{B}_{n+1}}Z \ \ell n \frac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z}\, dP \geq 0$$

(recall that $E^{\mathcal{B}_{n+1}}(E^{\mathcal{B}}Z) = E^{\mathcal{B}_{n+1}}Z$ and $E^{\mathcal{B}_{n+1}}(E^{\mathcal{B}_n}Z) = E^{\mathcal{B}_n}Z)$ we may write (3.1) as

$$(3.3) \quad \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_n}Z}\, dP = \int E^{\mathcal{B}_{n+1}}Z \ \ell n \frac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z}\, dP +$$

$$= \int E^{\mathcal{B}}Z \ \ell n \frac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_{n+1}}Z}\, dP.$$

Hence using the notation

(3.4) $I(\mathcal{B}:\mathcal{B}_n) = \int E^{\mathcal{B}}Z \; \ell n \dfrac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_n}Z} \, dP$

we have the chain of inequalities

(3.5) $I(\mathcal{B}:\mathcal{B}_0) \geq I(\mathcal{B}:\mathcal{B}_1) \geq \ldots \geq \quad I(\mathcal{B}:\mathcal{B}_n) \geq I(\mathcal{B}:\mathcal{B}_{n+1}) \geq \ldots \geq 0$

where it is seen from (3.3) that $I(\mathcal{B}:\mathcal{B}_n) = I(\mathcal{B}:\mathcal{B}_{n+1}) \Leftrightarrow$

$\Leftrightarrow I(\mathcal{B}_{n+1}:\mathcal{B}_n) = 0$. Since the monotonically decreasing sequence (3.5) is bounded below it converges and hence

(3.6) $I(\mathcal{B}:\mathcal{B}_n) - I(\mathcal{B}:\mathcal{B}_{n+1}) = \int E^{\mathcal{B}_{n+1}}Z \; \ell n \dfrac{E^{\mathcal{B}_{n+1}}Z}{E^{\mathcal{B}_n}Z} \, dP \to 0, \; n \to \infty.$

Hence using the result in [12], (3.6) implies

(3.7) $\int |E^{\mathcal{B}_{n+1}}Z - E^{\mathcal{B}_n}Z| \, dP \to 0, \; n \to \infty.$

Indeed, by considering $I(\mathcal{B}:\mathcal{B}_{n+m}) - I(\mathcal{B}:\mathcal{B}_n)$ we can also get

(3.8) $\int |E^{\mathcal{B}_{n+m}}Z - E^{\mathcal{B}_n}Z| \, dP \to 0 \quad m,n \to \infty,$

hence the sequence $E^{\mathcal{B}_n}Z$ is L_1 fundamental and there exists an $X \in L_1$ such that

(3.9) $\int |E^{\mathcal{B}_n}Z - X| \, dP \to 0, \; n \to \infty,$

and

(3.10) $\int E^{\mathcal{B}_n} Z dP \to \int X dP, \quad n \to \infty,$

that is

(3.11) $\int X dP = 1$

since $\int E^{\mathcal{B}_n} Z dP = 1$ for all n (see [13, p. 157, 161]).

Without restricting the generality we may take X to be $\mathcal{B}$-measurable by the following argument. Using the result in [13, p. 348] that

$$X_n \xrightarrow{L_1} X \Rightarrow E^{\mathcal{B}} X_n \xrightarrow{L_1} E^{\mathcal{B}} X$$

we have

$$\int |E^{\mathcal{B}_n} Z - X| dP \to 0, \; n \to \infty \;\; \Rightarrow \int |E^{\mathcal{B}} E^{\mathcal{B}_n} Z - E^{\mathcal{B}} X| dP \to 0,$$

$n \to \infty$

$$\Rightarrow \int |E^{\mathcal{B}_n} Z - E^{\mathcal{B}} X| dP \to 0, \; n \to \infty$$

since $E^{\mathcal{B}} E^{\mathcal{B}_n} Z = E^{\mathcal{B}_n} Z$. If X is $\mathcal{B}$-measurable $E^{\mathcal{B}} X = X$ a.s.

We shall now show that $X = E^{\mathcal{B}} Z$ a.s. Applying lemma 0.1 to (3.9) we have

(3.12) $\int_A E^{\mathcal{B}_n} ZdP \to \int_A XdP.$

If $A \in \mathcal{B}_k$, then for $n \geq k$

(3.13) $\int_A E^{\mathcal{B}_n} ZdP = \int_A E^{\mathcal{B}} ZdP$

so that

(3.14) $\int_A XdP = \int_A E^{\mathcal{B}} ZdP.$

Since (3.14) is true for $A \in \mathcal{B}_k$, it is true for $A \in \cup \mathcal{B}_k$. Thus the probability measures defined by the integrals in (3.14) are identical on the field $\bigcup\limits_k \mathcal{B}_k$ and hence by the extension theorem [13, p. 87] (3.14) holds for $A \in \mathcal{B}$. The Radon-Nikodym theorem then yields from (3.14) that

(3.15) $X = E^{\mathcal{B}} Z [P_{\mathcal{B}}].$

Since

(3.16) $\int |E^{\mathcal{B}_n} Z - E^{\mathcal{B}} Z| dP \leq \int |E^{\mathcal{B}_n} Z - X| dP + \int |X - E^{\mathcal{B}} Z| dP$

we see from (3.9) and (3.15) that (cf. [4, pp. 319, 331])

(3.17) $\int |E^{\mathcal{B}_n} Z - E^{\mathcal{B}} Z| dP \to 0, \quad n \to \infty.$

Since, using theorem 2.2

(3.18) $\int E^{\mathcal{B}} Z \, \ln \frac{E^{\mathcal{B}} Z}{X} \, dP = 0$

if and only if $X = E^{\mathcal{B}} Z$ a.s., then in view of (3.15) we conjecture that the sequence (3.5) has the limit zero, that is,

(3.19) $\lim\limits_{n \to \infty} I(\mathcal{B} : \mathcal{B}_n) = 0$

We show that (3.19) is true as follows. We write

$$I(\mathcal{B}:\mathcal{B}_n) = \int E^{\mathcal{B}}Z \,\ln \frac{E^{\mathcal{B}}Z}{E^{\mathcal{B}_n}Z}\, dP = \int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}}Z\, dP - \int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}_n}Z\, dP$$

$$= \int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}}Z\, dP - \int E^{\mathcal{B}_n}Z \,\ln E^{\mathcal{B}_n}Z\, dP.$$

We have shown that $E^{\mathcal{B}_n}Z \overset{L_1}{\to} E^{\mathcal{B}}Z$ which implies that $E^{\mathcal{B}_n}Z \overset{P}{\to} E^{\mathcal{B}}Z$.

The convergence in probability implies that there exists a sequence $\{n_k\}$ of integers increasing to infinity such that [13, p. 151]

$$E^{\mathcal{B}_{n_k}}Z \overset{a.s.}{\to} E^{\mathcal{B}}Z.$$

Since the convex function $E^{\mathcal{B}_n}Z \,\ln E^{\mathcal{B}_n}Z \geq -1/e$, the Fatou-Lebesgue Theorem [13, p. 152] yields

(3.20) $\liminf \int E^{\mathcal{B}_{n_k}}Z \,\ln E^{\mathcal{B}_{n_k}}Z\, dP \geqq \int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}}Z\, dP.$

But the convexity, Jensen's inequality, and the smoothing property of conditioning [13, p. 348, 351] lead to

(3.21) $\int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}}Z\, dP \geqq \int E^{\mathcal{B}_n}Z \,\ln E^{\mathcal{B}_n}Z\, dP \geqq \int E^{\mathcal{B}_m}Z \,\ln E^{\mathcal{B}_m}Z\, dP$

for all n, m such that n>m.

From the monotonic property in (3.21) combined with (3.20) we conclude that

$$\lim \int E^{\mathcal{B}_{n_k}}Z \,\ln E^{\mathcal{B}_{n_k}}Z\, dP = \int E^{\mathcal{B}}Z \,\ln E^{\mathcal{B}}Z\, dP$$

or

$$\lim I(\mathcal{B}:\mathcal{B}_{n_k}) = 0.$$

Since $I(\mathcal{B}:\mathcal{B}_n)$ converges, it must converge to the same limit, that is, we have (3.19).

Note that if we interpret $I(\mathcal{B}:\mathcal{B}_n)$ as a measure of the closeness of the approximation to the measures over $\mathcal{B}$ by the measures over $\mathcal{B}_n$ the sequence (3.5) implies that the approximation gets better as n gets larger and in the limit the approximation is exact to within sets of measure zero.

A similar argument of course follows for Z_t. If we start with

$$(3.22)\quad \int E_Z^{\mathcal{B}} W_t \;\ell n \frac{E_Z^{\mathcal{B}} W_t}{E_Z^{\mathcal{B}_n} W_t}\, d\mu = \int E_Z^{\mathcal{B}} W_t \;\ell n \frac{E_Z^{\mathcal{B}_{n+1}} W_t}{E_Z^{\mathcal{B}_n} W_t}\, d\mu \; + $$

$$+ \int E_Z^{\mathcal{B}} W_t \;\ell n \frac{E_Z^{\mathcal{B}} W_t}{E_Z^{\mathcal{B}_{n+1}} W_t}\, d\mu$$

since

$$(3.23)\quad \int E_Z^{\mathcal{B}} W_t \;\ell n \frac{E_Z^{\mathcal{B}_{n+1}} W_t}{E_Z^{\mathcal{B}_n} W_t}\, d\mu = \int E_Z^{\mathcal{B}_{n+1}} W_t \;\ell n \frac{E_Z^{\mathcal{B}_{n+1}} W_t}{E_Z^{\mathcal{B}_n} W_t}\, d\mu$$

we can repeat the preceding argument and conclude that

$$(3.24)\quad \begin{cases} \int |E_Z^{\mathcal{B}_n} W_t - Y_t|\, d\mu \to 0, \quad n \to \infty \\ \int Y_t\, d\mu = 1, \\ Y_t = E_Z^{\mathcal{B}} W_t \quad [\mu_{\mathcal{B}}] \end{cases}$$

$$(3.25)\quad \lim_{n\to\infty} \int E_Z^{\mathcal{B}} W_t \ \ln \frac{E_Z^{\mathcal{B}} W_t}{E_Z^{\mathcal{B}_n} W_t}\, d\mu = 0.$$

We shall now use these results to show that

$$(3.26)\quad I(\mathcal{B}) = \int E_Z^{\mathcal{B}} W_t \ \ln E_Z^{\mathcal{B}} W_t \ d\mu = \lim_{n\to\infty} \int E_Z^{\mathcal{B}_n} W_t \ \ln E_Z^{\mathcal{B}_n} W_t \ d\mu =$$

$$= \lim_{n\to\infty} I(\mathcal{B}_n).$$

Using the result in theorem 2.3 we may write

$$(3.27)\quad \begin{aligned} I(\mathcal{B}|\mathcal{B}_0) &= I(\mathcal{B}_1|\mathcal{B}_0) + I(\mathcal{B}|\mathcal{B}_1) \\ I(\mathcal{B}|\mathcal{B}_1) &= I(\mathcal{B}_2|\mathcal{B}_1) + I(\mathcal{B}|\mathcal{B}_2) \\ \ldots \quad & \quad \ldots \quad \ldots \\ I(\mathcal{B}|\mathcal{B}_k) &= I(\mathcal{B}_{k+1}|\mathcal{B}_k) + I(\mathcal{B}|\mathcal{B}_{k+1}) \\ \ldots \quad & \quad \ldots \quad \ldots \end{aligned}$$

or using the relations in (3.27)

$$(3.28)\quad I(\mathcal{B}|\mathcal{B}_0) = I(\mathcal{B}) = \sum_{k=0}^{n} I(\mathcal{B}_{k+1}|\mathcal{B}_k) + I(\mathcal{B}|\mathcal{B}_{n+1})$$

where

$$(3.29)\quad I(\mathcal{B}|\mathcal{B}_N) = \int E_Z^{\mathcal{B}} W_t \ \ln \frac{E_Z^{\mathcal{B}} W_t}{E_Z^{\mathcal{B}_N} W_t}\, d\mu.$$

Since the relation (3.25) applied to (3.29) implies

$$(3.30)\quad I(\mathcal{B}|\mathcal{B}_N) \to 0 \text{ as } N \to \infty$$

we get from (3.28) (see (2.8), (2.10))

$$(3.31)\quad I(\mathcal{B}) = \sum_{k=0}^{\infty} I(\mathcal{B}_{k+1}|\mathcal{B}_k) = \sum_{k=0}^{\infty} (I(\mathcal{B}_{k+1}) - I(\mathcal{B}_k)) = \lim_{n\to\infty} I(\mathcal{B}_n).$$

From (3.31) it is seen that if $I(\mathcal{B}) < \infty$ then $I(\mathcal{B}_{n+1}|\mathcal{B}_n) \to 0$,

$n \to \infty$. If $I(B_{n+1} | B_n) \to I < \infty$ as $n \to \infty$, then by the Toeplitz lemma [13, p. 238] $\lim_{n \to \infty} \frac{1}{n} I(B_n) = I$.

We are now in a position to apply (3.31) to prove the integral representation theorem.

In the preliminaries it was seen that a separable sigma-algebra can be generated by a regular sequence of partitions, and using the notation of (0.18) we have

$$(3.32) \quad \underline{E}^n \uparrow \underline{E} \; .$$

Accordingly, as a special case of (3.31) we have that

$$(3.33) \quad \lim_{n \to \infty} I(\underline{E}^n) = I(\underline{E}) \quad .$$

Recalling (1.11), (2.15) and Theorem 2.5, we note that (3.33) is the integral representation theorem, that is

$$(3.34) \quad I(A) = \operatorname{Sup} \sum \mu_t(A_i) \, \ln \frac{\mu_t(A_i)}{\mu(A_i)} = \bar{I}(A)$$

where the sup is taken over all possible A-measurable finite partitions of Ω, and in particular if the B_n are generated by finite partitions then (3.31) is the same as (2.4).

For application of the preceding results, particularly (3.31) to stochastic processes we state as Lemma 3.1 a result which is Theorem 1.6, page 604 of [4].

<u>Lemma 3.1</u>. Let A be a sigma-algebra of ω sets, and let $\{x(t,\omega), t \in T\}$ be a family of ω functions measurable with

respect to Λ. Let $\mathcal{B}_S$ be the sigma-algebra generated by $\{x(t,\omega),\ t\in S\subset T\}$. Suppose that T is non-denumerable. Then if $\Lambda\in\mathcal{B}_T$ there is a denumerable subset S (depending on Λ) of T, such that $\Lambda\in\mathcal{B}_S$. If $X(\omega)$ is an ω function measurable with respect to $\mathcal{B}_T$, there is a denumerable subset S (depending on X) such that X is measurable with respect to $\mathcal{B}_S$.

Now let $\{x(t,\omega),\ t\in T\}$ be an arbitrary system of random variables defining a stochastic process. Let $\mathcal{B}_N$ be the sigma-algebra generated by the sub-system $\{x(t,\omega),\ t\in N\subset T\}$ and $\mathcal{B}_T$ the sigma-algebra generated by the system of random variables defining the stochastic process. We can now state

Theorem 3.1.

$$I(\mathcal{B}_T) = \sup_{N\in\mathcal{N}} I(\mathcal{B}_N) \tag{3.35}$$

where $\mathcal{N}$ is the class of all finite subsets of T.

Proof. If T is countable, then it is possible to choose finite subsets $N_1 \subset N_2 \subset \dots$ such that $\mathcal{B}_T$ is the smallest sigma-algebra containing $\bigcup_{i=1}^{\infty} \mathcal{B}_{N_i}$ and (3.35) is then essentially (3.31). If T is not countable we shall use Lemma 3.1. If $I(\mathcal{B}_S) = \infty$, then $I(\mathcal{B}_T) = \infty$. Suppose $I(\mathcal{B}_T) < \infty$, then $E_Z^{\mathcal{B}_T} W_t = E^{\mathcal{B}_T} Z_t / E^{\mathcal{B}_T} Z$ (see (0.5)) exists and is of course

measurable with respect to $\mathcal{B}_T$. But according to Lemma 3.1 every function measurable with respect to $\mathcal{B}_T$ is measurable with respect to $\mathcal{B}_S$ for at least one countable $S \subset T$ so that

$$(3.36)\quad E_Z^{\mathcal{B}_T} W_t = E_Z^{\mathcal{B}_S} W_t \ [\mu], \text{ or } E^{\mathcal{B}_T} Z_t / E^{\mathcal{B}_T} Z = E^{\mathcal{B}_S} Z_t / E^{\mathcal{B}_S} Z \quad \text{a.s.}$$

Since (3.36) is the necessary and sufficient condition that $I(\mathcal{B}_S) = I(\mathcal{B}_T)$ we get

$$(3.37)\quad I(\mathcal{B}_T) = I(\mathcal{B}_S) = \sup_{N \in \mathcal{N}} I(\mathcal{B}_N).$$

For a similar result see [9].

4. Monotonicity. If the sigma-algebra $\mathcal{B}$ in (3.0) is not the sigma-algebra $\mathcal{A}$ of the probability space $(\Omega, \mathcal{A}, P)$ then again using theorem 2.3 as in (3.27) we have

$$(4.1)\quad I(\mathcal{A}) = I(\mathcal{B}) + I(\mathcal{A}|\mathcal{B})$$

$$(4.2)\quad I(\mathcal{A}) = I(\mathcal{B}_n) + I(\mathcal{A}|\mathcal{B}_n).$$

We can now derive certain limiting relations in which $\mathcal{A}$ plays a role. From (4.2) and (3.31) we see that

$$(4.3)\quad I(\mathcal{A}) = \lim_{n\to\infty} I(\mathcal{B}_n) + \lim_{n\to\infty} I(\mathcal{A}|\mathcal{B}_n)$$

$$= I(\mathcal{B}) + \lim_{n\to\infty} I(\mathcal{A}|\mathcal{B}_n),$$

hence for $I(\mathcal{A}) < \infty$

$$(4.4)\quad I(\mathcal{A}) - I(\mathcal{B}) = I(\mathcal{A}|\mathcal{B}) = \lim_{n\to\infty} I(\mathcal{A}|\mathcal{B}_n).$$

Similarly from

$$(4.5)\quad I(\mathcal{B}) = I(\mathcal{B}_n) + I(\mathcal{B}|\mathcal{B}_n)$$

(4.6) $I(B) = \lim_{n\to\infty} I(B_n) + \lim_{n\to\infty} I(B|B_n)$

so that if $I(B) < \infty$

(4.7) $I(B) = \lim_{n\to\infty} I(B_n) \Leftrightarrow \lim_{n\to\infty} I(B|B_n) = 0.$

Note that if $I(A) < \infty$ and $B_m \subset B_n$, then using (4.2) we have

(4.8) $I(A|B_m) - I(A|B_n) = I(B_n) - I(B_m) \geq 0.$

Similarly for $B_k \subset B_m \subset B_n$ we have

(4.9) $I(B_n|B_k) \geq I(B_m|B_k).$

As a matter of fact for $B_1 \subset B_2 \subset A$ we can write if $I(A) < \infty$ (for a related discussion see [8])

(4.10) $I(A|B_1) = I(B_2|B_1) + I(A|B_2)$

which can be proven either directly or from the fact that

(4.11) $[I(A) - I(B_1)] = [I(B_2) - I(B_1)] + [I(A) - I(B_2)].$

Since the information values in (4.10) are nonnegative it follows that

(4.12) $I(A|B_1) \geq I(A|B_2)$

with equality if and only if $I(B_2|B_1) = 0$ and

(4.13) $I(A|B_1) \geq I(B_2|B_1)$.

with equality if and only if $I(A|B_2) = 0.$

Using A in place of B in (3.27) we have corresponding to (3.28)

(4.14) $I(A|B_0) = I(A) = \sum_{k=0}^{n} I(B_{k+1}|B_k) + I(A|B_{k+1})$

so that using (4.4)

$$(4.15) \quad I(A) = \sum_{k=0}^{\infty} I(B_{k+1}|B_k) + I(A|B).$$

In connection with the results of this section see the axiomatic approach in [5].

REFERENCES

[1] Chiang Tse-Pei (1958). A note on the definition of information. Theory of Probability and Its Applications (English translation) 3, 93-97.

[2] Dobrushin, R. L. (1959). General formulation of Shannon's main theorem in information theory. (Russian) Usp. Mat. Nauk., 14, No. 6, 3-104. English translation: A.M.S. Translations, Series 2, 33 (1963), 323-438.

[3] Dobrushin, R. L. (1960). Passage to the limit under the information and entropy signs. Theory of Probability and Its Applications. (English translation) 5, 25-32.

[4] Doob, J. L. (1953). Stochastic Processes, Wiley, New York.

[5] Forte, B. (1970). Generalized measures of information and uncertainty. Proceedings of the Meeting on Information Measures, University of Waterloo, Ontario, Canada, April 10-14, 1970.

[6] Gelfand, I. M., Kolmogorov, A. N., and Yaglom, A. M. (1956). On the general definition of the amount of information (Russian), Doklady Akad, Nauk, SSSR., 111, 745-748. German translation: Arbeiten zur Informations theorie II, VEB Deutscher Verlag der Wissenschaften, Berlin, 1958, 57-60.

[7] Gelfand, I. M. and Yaglom, A. M. (1957). Calculation of the amount of information about a random function contained in another such function. (Russian) Usp. Mat. Nauk., 12, No. 1 (73), 3-52. English translation: A.M.S. Translations, Series 2, 12 (1959), 199-246.

[8] Ghurye, S. G. (1968). Information and sufficient subfields. Annals of Math. Statist. 39, 2056-66.

[9] Hájek, J. (1958). A property of J-divergence of marginal probability distributions. Czechoslovak Math. J. 8 (83), 460-3.

[10] Kallianpur, G. (1960). On the amount of information contained in a σ-field. Contributions to Probability and Statistics-Essays in Honor of Harold Hotelling. Stanford University Press, 265-73.

[11] Kullback, S. (1959). Information Theory and Statistics. Wiley, New York, 1968 Edition, Dover Publications Inc., New York.

[12] Kullback, S. (1967). A lower bound for discrimination information in terms of variation. IEEE Transactions on Information Theory, IT-13, 126-7.

[13] Loéve, M. (1955). Probability Theory. D. Van Nostrand Co., Inc., Second Edition, 1960; Third Edition, 1963.

[14] Pérez, A. (1957). Notions generalisees d'incertitude d'entropie et d'information du point de vue de la theorie de martingales. Transactions of the first Prague conference on information theory, statistical decision functions, random processes., November 28 to 30, 1956. Publishing house of the Czechoslovak Academy of Sciences, Prague, 183-208.

[15] Pinsker, M. S. (1964). Information and Information Stability of Random Variables and Processes, Holden-Day Inc. (translated by A. Feinstein from the Russian published in Moscow, 1960).